AI Transparency in the Age of LLMs: A Human-Centered Research Roadmap

Year: 2023 | Citations: 216 | Authors: Q. Liao, Jennifer Wortman Vaughan

---

Abstract

The rise of powerful large language models (LLMs) brings about tremendous opportunities for innovation but also looming risks for individuals and society at large. We have reached a pivotal moment for ensuring that LLMs and LLM-infused applications are developed and deployed responsibly. However, a central pillar of responsible AI -- transparency -- is largely missing from the current discourse around LLMs. It is paramount to pursue new approaches to provide transparency for LLMs, and years of research at the intersection of AI and human-computer interaction (HCI) highlight that we must do so with a human-centered perspective: Transparency is fundamentally about supporting appropriate human understanding, and this understanding is sought by different stakeholders with different goals in different contexts. In this new era of LLMs, we must develop and design approaches to transparency by considering the needs of stakeholders in the emerging LLM ecosystem, the novel types of LLM-infused applications being built, and the new usage patterns and challenges around LLMs, all while building on lessons learned about how people process, interact with, and make use of information. We reflect on the unique challenges that arise in providing transparency for LLMs, along with lessons learned from HCI and responsible AI research that has taken a human-centered perspective on AI transparency. We then lay out four common approaches that the community has taken to achieve transparency -- model reporting, publishing evaluation results, providing explanations, and communicating uncertainty -- and call out open questions around how these approaches may or may not be applied to LLMs. We hope this provides a starting point for discussion and a useful roadmap for future research.